View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Springer - Publisher Connector

BMC Bioinformatics

BioMed Central

Research article

Prediction of MHC class I binding peptides, using SVMHC

Pierre Dönnes[1] and Arne Elofsson*[2]

Address: [1]Center for Bioinformatics Saar, Saarland University, D-660 41 Saarbrücken, Germany and [2]Stockholm Bioinformatics Center, SCFAB, Stockholm University, SE-106 91 Stockholm, Sweden

E-mail: Pierre Dönnes - pierre@bioinf.uni-sb.de; Arne Elofsson* - arne@sbc.su.se

*Corresponding author

Published: 11 September 2002

BMC Bioinformatics 2002, **3**:25

Received: 22 March 2002

Accepted: 11 September 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/25

© 2002 Dönnes and Elofsson; licensee BioMed Central Ltd. This article is published in Open Access: verbatim copying and redistribution of this article are permitted in all media for any non-commercial purpose, provided this notice is preserved along with the article's original URL.

Keywords: MHC class I, Peptide prediction, Machine Learning, Support Vector Machines

Abstract

Background: T-cells are key players in regulating a specific immune response. Activation of cytotoxic T-cells requires recognition of specific peptides bound to Major Histocompatibility Complex (MHC) class I molecules. MHC-peptide complexes are potential tools for diagnosis and treatment of pathogens and cancer, as well as for the development of peptide vaccines. Only one in 100 to 200 potential binders actually binds to a certain MHC molecule, therefore a good prediction method for MHC class I binding peptides can reduce the number of candidate binders that need to be synthesized and tested.

Results: Here, we present a novel approach, SVMHC, based on support vector machines to predict the binding of peptides to MHC class I molecules. This method seems to perform slightly better than two profile based methods, SYFPEITHI and HLA_BIND. The implementation of SVMHC is quite simple and does not involve any manual steps, therefore as more data become available it is trivial to provide prediction for more MHC types. SVMHC currently contains prediction for 26 MHC class I types from the MHCPEP database or alternatively 6 MHC class I types from the higher quality SYFPEITHI database. The prediction models for these MHC types are implemented in a public web service available at [http://www.sbc.su.se/svmhc/].

Conclusions: Prediction of MHC class I binding peptides using Support Vector Machines, shows high performance and is easy to apply to a large number of MHC class I types. As more peptide data are put into MHC databases, SVMHC can easily be updated to give prediction for additional MHC class I types. We suggest that the number of binding peptides needed for SVM training is at least 20 sequences.

Background

As the genome projects proceed, we are presented with an exponentially increasing number of known protein sequences. Sequences from pathogens provide a huge amount of potential vaccine candidates, as the activation of cytotoxic T-cells requires recognition of specific peptides bound to Major Histocompatibility Complex (MHC) class I molecules (for humans the term Human Leukocyte Antigens, HLA, is often used instead of MHC). MHC-binding peptides (MHC-peptides) are also potential tools for diagnosis and treatment of cancer [1]. However, it is estimated that only one in 100 to 200 peptides

actually binds to a particular MHC [2]. Therefore, a good computational prediction method could significantly reduce the number of peptides that have to be synthesized and tested.

Prediction of MHC-peptides can be divided into two groups: sequence based and structure based methods. Allele specific sequence motifs can be identified by studying the frequencies of amino acids in different positions of identified MHC-peptides. The peptides that bind to HLA-A*0201 are often 9 amino acids long (nonamers), and frequently have two anchor residues, a lysine in position 2 and a Valine in position 9 [3]. This type of sequence patterns has been used as a simple prediction method [4]. Besides the anchor residues, there are also weaker preferences for specific amino acids in other positions. One method to include this information is to use a profile, where a score is given for each type of amino acid in each position [5]. The scores can be calculated from observed amino-acid frequencies in each position or be set manually. The sum of the scores for a given peptide is then used to make predictions. One frequently used profile based prediction method is SYFPEITHI [6], which is freely available as a web service at [http://www.syfpeithi.de/]. The matrices in SYFPEITHI were adjusted manually, by assigning a high score (10) for frequently occurring anchor residues, a score of 8 to amino acids that occur in a significant amount and a score of 6 to rarely occurring residues. Preferred amino acids in other positions have scores that range from 1 to 6 and amino acids regarded as unfavorable have scores ranging from -3 to -1. SYFPEITHI prediction can be done for 13 different MHC class I types. Another profile based MHC-peptide predictor is HLA_BIND at [http://bimas.dcrt.nih.gov/molbio/hla_bind/]. This method estimates the half-time of dissociation of a given MHC-peptide complex [7]. HLA_BIND provides prediction for more than 40 different MHC class I types. It has been shown that profile based methods are correct in about 30% of the time, in the sense that one third of the predicted binders actually bind [8].

A profile based method does not take into account correlations between frequencies in different positions, neither they consider information from peptides that do not bind. This information can be used by machine learning methods [9]. Prediction of MHC-peptides has been made by using machine learning approaches such as artificial neural networks [10] and hidden Markov models [11]. Gulukota et al. (1997) [8] showed that one advantage of machine learning algorithms compared to profile methods seems to be that they have a higher specificity. This is possible due to the inclusion of non-binding data in the training. A machine learning approach extracts useful information from a large amount of data and creates a good probabilistic model [9]. In the case of MHC-peptide prediction, a data set of known binders and known (or supposed) non-binders is used. This set is then used to build a model that discriminates between binding peptides and non-binding peptides. This model can then be used to predict whether a novel peptide binds or not. Brusic reported a total accuracy of 88% on predictions for the mouse MHC H-2K^d, using an artificial neural networks and hidden Markov models have been reported to perform 2–15% better than artificial neural networks [11,12].

Structural approaches for prediction evaluate how well a peptide fit in the binding groove of a MHC molecule. A peptide is threaded through a structural template to obtain a rough estimate of the binding energy. The energy estimation is based on the interactions defined in the binding pocket of a particular MHC molecule [13]. To our knowledge no comparisons of the performance between structural and sequence based methods has been published. Obviously, a structural approach is limited to MHC types with a known structure. However, the advantage of a structural approach is that one known structure alone might be sufficient for creating a prediction model.

Results and Discussion

The amount of known binding data for different MHC molecules varies significantly. For some MHC molecules only a few MHC-peptides are known, while for others, there are several hundred verified binders. Since all machine learning methods need a sufficient amount of data for training, we investigated the number of known binders needed for training, using three examples with a large set of known binders. A varying number of training examples was tested using the nonamers in MHCPEP binding to HLA-A*0201, HLA-A3 and HLA-B*2705. The ratio of positive/negative examples was kept constant at 1:2. The test sets for each of the three HLA types consisted of 20 binders and 40 non-binders, unrelated from the training sets. The Mc for the test set was calculated for each size of the training set. A significant improvement of Mc was observed when the size of the training set was increased up to about 20 MHC-peptides, see figure 1. Further, a smaller improvement was observed for up to 50 peptides. From the similar behavior of these three examples we concluded that it seemed necessary to include at least 20 known peptides for successful predictions. This resulted in that the current version of SVMHC can make predictions for 26 different MHC molecules, using MHCPEP data. If SYFPEITHI data was used prediction could only be done for 6 different MHC molecules.

The overall performance of SVMHC was compared to SYFPEITHI and HLA_BIND for the six MHC types common between the methods. In Table 1 it can be seen that SVMHC in general performs slightly better than SYFPEITHI

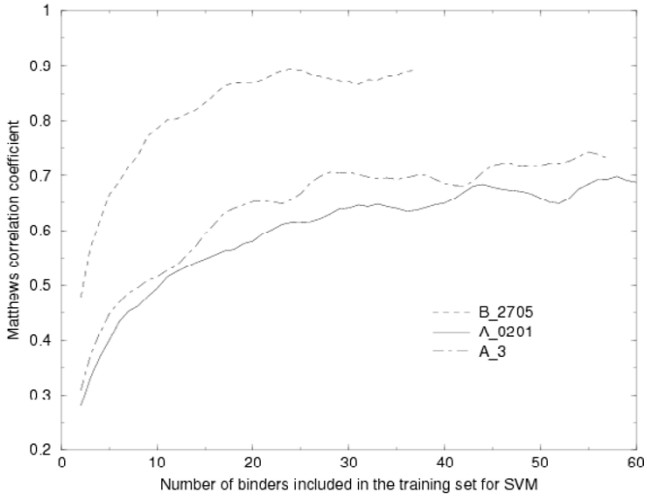

Figure 1
Performance of SVMHC for the MHC type HLA-A*0201, HLA-A3 and HLA-B*2705, measured by the Matthews correlation coefficient, Mc, versus the number of peptides used for training. For all sizes of the training-set the test-set was identical and no part of the test-set was contained in the training-set.

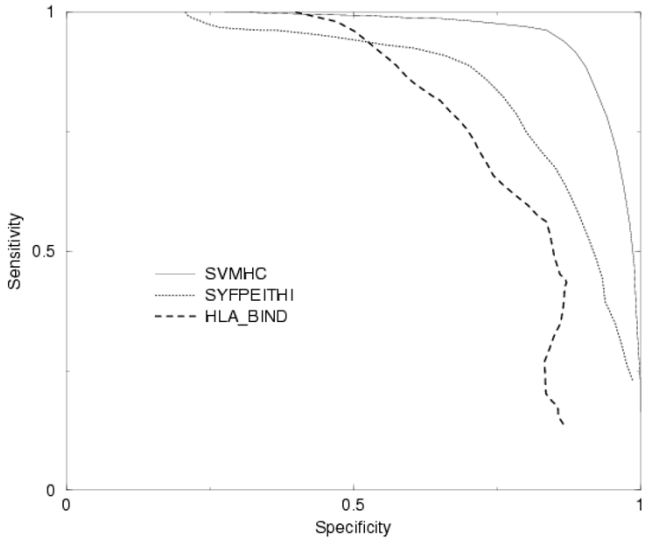

Figure 2
Specificity/sensitivity plots for SVMHC HLA_BIND and SYFPEITHI. Sensitivity is defined as the number of correctly predicted binders (TP) found at a given cutoff, divided with the total number of binders, i.e. Sens = TP/(TP+FN), where FN is the number of . The specificity is defined as the fraction of the hits above this cutoff that is correct, i.e. Spec = TP/(TP+FP). It can be seen that the sensitivity of SVMHC is higher than of SYFPEITHI and HLA_BIND at any specificity.

and HLA_BIND. SVMHC correctly identified 95% of the peptides, while SYFPEITHI and HLA_BIND only classified 91% and 87% of the peptides correctly. In figure 2, it can also be seen that 90% of all MHC-peptides can be identified at a specificity of 90% by SVMHC, while this sensitivity is only reached at a specificity of 75% by SYFPEITHI and at 50% for HLA_BIND. It seems as if the hand-tuned profiles from SYFPEITHI performs slightly better than HLA_BIND and in agreement with earlier studies machine learning based methods, such as SVMHC, show a higher specificity than profile based methods [8]. When studying the performance of the individual alleles it can be seen that in five of these cases SVMHC show the best performance, only for HLA-B*8 HLA_BIND performs slightly better. HLA-B*8 is also the allele with the lowest number of known binders.

In addition to the overall performance we have studied the performance of all MHC classes in SVMHC, see table 3. It can be seen that the prediction quality varies between MC = 0.59 and 1.0. The worse predictions are for two datasets with few data-points, decamers for HLA-A2 and HLA-A11.

Finally, we tested SVMHC by performing a prediction for four proteins with recently identified known MHC-peptides. All possible binding nonamers were run through the predictors and a ranked list of candidate binders was produced from the output (the SVMHC models used were trained on SYFPEITHI data). In table 2 it can be seen that for all four proteins SVMHC ranks the known binders higher than the other two methods. This also indicates that fewer non-binders are given high scores when using SVMHC.

This example further supports the suggestion that machine learning methods might improve the specificity over profile based methods. However, the increase over SYFPEITHI, seems quite marginal and the major advantage of SVMHC might be that it (a) contains more MHC-types, (b) the scores are comparable between different MHC types (c) a slightly higher specificity.

Conclusions

Here, we present a novel approach based on support vector machines to predict the binding of peptides to MHC class I molecules, SVMHC. This method seems to perform slightly better than profile based methods. Most importantly the scoring is more comparable between different MHC types and therefore provides a higher overall specificity. Moreover, the implementation of SVMHC was done in such way that so that it will be easy to update when new binding peptides are identified. Better methods for purification and sequencing of MHC-binding peptides are developed all the time, giving more accurate databases. Therefore, the use of more "high quality" data will increase the performance of SVMHC prediction in the future, and predictions of a larger number of MHC-I classes

Table 1: Comparison between SVMHC, SYFPEITHI (SYF) and HLA_BIND (HLA) of the six alleles common between them.

Dataset			Mc			percentage correct predictions		
MHC	Length	Size	SVMHC	SYF	HLA	SVMHC	SYF	HLA
Overall	-	-	0.85	0.75	0.62	95%	91%	87%
HLA-A*0201	9	113	0.78	0.77	0.77	90%	89%	89%
HLA-A*0201	10	40	0.70	0.61	0.61	87%	80%	83%
HLA-A1	9	28	0.96	0.93	0.96	98%	97%	98%
HLA-A3	9	73	0.80	0.73	0.71	91%	86%	84%
HLA-B*8	9	25	0.79	0.79	0.82	91%	91%	92%
HLA-B*2705	9	29	1.00	0.92	0.93	100%	95%	97%

The tables shows the MHC-type, the length of the binding peptides, the number of experimentally verified binders, the Matthew correlation coefficient (Mc) and the percentage correct predictions.

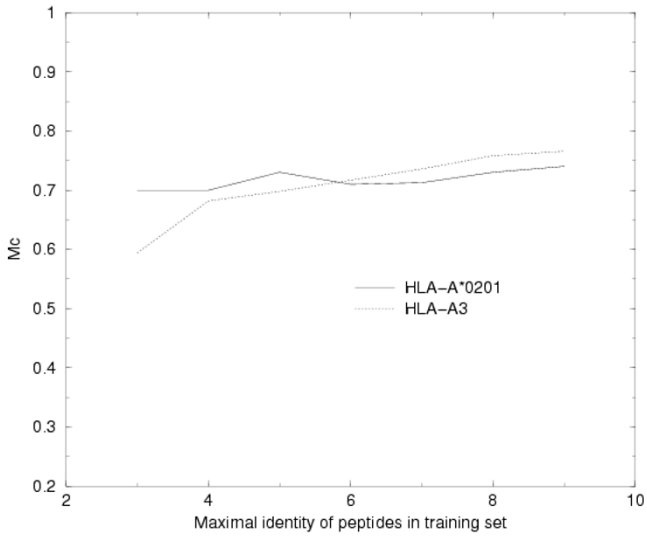

Figure 3
The dependency of Matthew correlation coefficient on the reduction level for two HLA alleles (HLA-A*0201 and HLA-A3). The reduction level is measured as the maximum number of allowed identical measures between two peptides in the set.

should also be available. SVMHC currently contains prediction models for 26 MHC class I types from the MHCPEP database and 6 MHC class I types from the SYFPEITHI database. The prediction models for these MHC types are implemented in a public web service available at [http://www.sbc.su.se/svmhc/].

Methods

This paper presents a support vector machine based method (SVMHC) to predict peptides that bind MHC class I molecules. Support vector machines are a class of machine learning methods that recently has been applied for classification of microarray data, protein structure prediction and other biological problems [14–16]. In preliminary studies, it was indicated that support vector machines performed better than neural networks for MHC-peptide predictions. SVMHC is based on the support vector machine package SVM-LIGHT [17].

Support vector machines

A full coverage of the use of SVM for pattern recognition is given by Vapnik [18], but some basic concepts are introduced here. Lets assume that we have a series of examples (or input vectors) $\bar{x}_i \in R^d$ (i = 1, 2 ..., N) with corresponding labels $y_i \in \{+1, -1\}$ (i = 1, 2, ..., N). In the case of MHC class I binding peptides, $\bar{x}_i$ corresponds to the amino acid sequence of the peptide and y_i (+1 or -1) represents binder/non-binder. The amino acid sequence of a peptide is represented by sparse encoding [9].

This task is carried out by (i) mapping of the input vectors $\bar{x}_i$ into a high dimensional feature space $\Phi(\bar{x}) \in H$ and (ii) construction of an optimal separating hyperplane (OSH) in the new feature space. The OSH is the hyperplane with the maximum distance to the nearest data points of each class in the feature space H. One of the most central points in using SVM is the choice of mapping $\phi(\cdot)$, which is defined by a kernel function $K(\bar{x}_i, \bar{x}_j)$. The decision function used by SVM can be written:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\bar{x}_i, \bar{x}_j + b)\right) \qquad (1)$$

The coefficients α_i are given by the solution of the quadratic programming task: Maximize

Table 2: Performance of SVMHC for different HLA alleles, using MHCPEP- or SYFPEITHI-data.

MHC	Length	Size	Mc	Kernel
Predictions using MHCPEP				
HLA-A1	9	28	0.95	lin
HLA-A*1101	9	40	0.74	poly
HLA-A11	9	46	0.75	rbf
HLA-A11	10	21	0.59	poly
HLA-A2	9	118	0.76	poly
HLA-A2	10	35	0.65	poly
HLA-A*2402	9	73	0.90	poly
HLA-A3	9	73	0.76	rbf
HLA-A*0201	9	184	0.73	rbf
HLA-A*0201	10	96	0.78	poly
HLA-A*3301	9	32	0.72	lin
HLA-A*0301	9	38	0.72	rbf
HLA-A*0301	10	32	0.77	lin
HLA-A31	9	39	0.79	poly
HLA-A*6801	9	42	0.84	poly
HLA-B7	9	32	0.95	lin
HLA-B8	9	26	0.77	poly
HLA-B*2705	9	41	0.93	lin
HLA-B*3501	9	67	0.93	lin
HLA-B*3501	10	34	0.96	poly
HLA-B35	9	23	0.71	lin
HLA-B*2703	9	22	0.90	lin
HLA-B*5301	9	41	0.95	lin
HLA-B27	9	34	0.91	rbf
HLA-B*2706	9	20	0.93	lin
HLA-B51	9	67	0.82	poly
HLA-B*5102	9	29	0.79	poly
HLA-B*0702	9	52	0.96	poly
HLA-B*5103	9	29	0.84	rbf
HLA-B*5401	9	42	0.98	lin
HLA-B*5101	9	35	0.89	lin
Predictions using SYFPEITHI				
HLA-A*0201	9	113	0.78	rbf
HLA-A*0201	10	40	0.70	poly
HLA-A1	9	28	0.96	lin
HLA-A3	9	73	0.80	lin
HLA-B*8	8	14	0.89	lin
HLA-B*8	9	25	0.79	lin
HLA-B*2705	9	29	1.00	lin
HLA-B7	9	23	0.93	lin

The first column explains shows the HLA allele, the second the length of the binding peptides, the third the number of binders included in the training set, the fourth the performance as measured by the Matthews correlation coefficient. The final column shows what type of kernel was used in the Support Vector Machine.

$$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j \cdot \gamma_i\gamma_j \cdot K\left(\overline{x}_i, \overline{x}_j\right)$$

subject to

$$0 \le \alpha_i \le c$$
$$\sum_{i=1}^{N}\alpha_i\gamma_i = 0, i = 1, 2, \ldots, N \quad (2)$$

c in equation (2) is a parameter controlling the trade off between the margin and and the training error. It is the kernel function that determines the dimension of the feature space, meaning that different kernels will represent the input vectors in different ways. The aim of the SVM is then to find an OSH without loosing the ability of generalization, often referred to as over-training. The kernels tested for MHC class I peptide predictions were linear, polynomial and radial basis function.

Equation (3) is an example of the radial basis function and equation (4) shows the polynomial kernel function.

$$K\left(\overline{x}_i, \overline{x}_j\right) = \exp\left(-\gamma\left\|\overline{x}_i - \overline{x}_j\right\|^2\right) \quad (3)$$
$$K\left(\overline{x}_i, \overline{x}_j\right) = \left(\overline{x}_i \bullet \overline{x}_j + 1\right)^d \quad (4)$$

The problem of choosing the most suitable kernel for a SVM is analogous to the problem of choosing the architecture for a neural network [15]. One main feature of SVM is that the quadratic programming task is a convex optimization problem, which ensure a global optimum. This can be compared to ANN that uses gradient based training functions with the risk of getting stuck in a local minimum.

SVMHC performance and parameter optimization

A central part of the process of developing a prediction method is to have a good measure of the prediction performance. The main goal is to have a prediction method that can generalize and correctly classify unseen data. Therefore, four-fold cross validation was used to verify SVM performance [19]. Further we have used redundancy reduction, such that no two peptides share more than four amino acids, see below. The main measure of performance used for SVMHC parameter optimization was Matthews Correlation coefficients (Mc) [20].

For each MHC-type the optimal kernel and trade off c was optimized by a systematic variation of the parameters and evaluation of prediction performance using Matthews Correlation coefficients. For the linear kernel the parameters j, a cost factor between errors on binding and non-binding peptides was also optimized. In the case of a polynomial and radial basis kernel the parameters describing the form of the function were optimized as well. The parameters chosen for each MHC class I type, were the ones that gave the best Matthews Correlation coefficient. For a more detailed explanation of the parameters, see the SVM-LIGHT documentation at [http://svmlight.joachims.org/].

MHC databases

In this study we have used two databases SYFPEITHI [6] and MHCPEP [21] to create MHC class I predictors for different alleles. MHCPEP is a curated database comprising over 13000 peptide sequences known to bind MHC molecules. Entries are compiled from published reports as well as from direct submissions of experimental data. SYFPEITHI is supposed to be of a higher quality and is restricted to published data and only contain sequences that are natural ligands to T-cell epitopes. The two databases have different advantages, MHCPEP contains signicantly more data (13000 vs 3500), while the quality of the data in SYFPEITHI is assumed to be higher. Therefore, using MHCPEP data for SVM training, it is possible to make predictions for 26 different MHC types. This can be compared with only 6 MHC types when SYFPEITHI data is used for SVM training. However, the predictions from SYFPEITHI might be more reliable and should therefore be used when enough data exists.

Peptide sequences known to bind a MHC class I alleles were extracted from one of the databases. All peptides from the two databases are considered as binding peptides, i.e. no difference between strong and weak binders is considered. Unfortunately, there are very few experimentally verified examples of peptides that do *not* bind to a particular MHC. Therefore, the non-binding training examples were extracted randomly from the ENSEMBL database of human proteins [22]. Protein sequences from the ENSEMBL database were chopped up into the length of interest and known MHC-peptides were removed. Obviously, there is a risk that some of the non-binders actually binds, but since less than 1% of the peptides are expected to bind to a MHC molecule, we do not expect this to cause any major problems. The ratio of binder/non-binders was kept to 1:2 for all MHC types.

Redundancy reduction

When utilizing machine learning methods it is important that the training data reproduces well what can be expected for unseen data. If the training data only contains a

Table 3: Study of recently detected binding peptides from proteins P17944, P17451, P31952 and P17944, binding to HLA-A*0201. The known binding nonamers and the rank of these with the different predictors are shown. Although all three methods detects these peptides, they are found at higher ranks using SVMHC than with the other methods.

Protein	No.	SVMHC	HLA_BIND	SYFPEITHI
P78395	2	1,6	3,23	2,8
P17944	2	1,2	1,3	4,6
P31952	1	2	4	14
P17451	3	2,3,4	1,4,12	1,6,10

subset of what can be expected there is a risk for overtraining, i.e. that the obtained performance is not representative for unseen data. One method to avoid overtraining is to use a "redundancy reduced" test-set. To understand the risk of over-training the data we used two alleles to study the change in performance using different reduction levels. We examined the performance on cross validated test-sets using different reduction levels for two different MHC alleles. For HLA-A*0201 the performance is not dependent on the reduction level, while a small increase is seen for HLA-A3 (from Mc = 0.60 to 0.74) when a looser cutoff is used, see figure 3. Using a stricter redundancy reduction might improve future predictions but as the dataset is limited it makes less alleles available for prediction. Therefore, in all studied below we choose to include a restriction that no two peptides in the dataset should share more than 4 identical residues.

Comparison of different prediction methods

The performance of SVMHC was compared to the performance of two public prediction servers, SYFPEITHI and HLA_BIND. The prediction performances were measured using Matthews Correlation coefficients (Mc) [20], Specificity-Sensitivity plots [23] and the percentage correct predictions. For SYFPEITHI and HLA_BIND the cutoff distinguishing between binders and non-binders was optimized, while for SVMHC it was kept constant and 0. There are six MHC types common between the three methods and all of these were used for comparing the performance. Each binding and non-binding peptide tested was submitted to the public prediction servers and the different prediction performances were calculated. The threshold for binder/non-binder for the public prediction servers, were chosen to give the maximum Mc on the test set.

Authors' contributions

PD carried out all of the work under the supervision of AE. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants form the Swedish Natural Sciences Research Council and the Swedish Research Council for Engineering Sciences to AE.

References

1. Sette A, Chesnut R, Fikes J: **HLA expression in cancer: implications for T cell-based immunotherapy.** *Immunogenetics* 2001, **53:**255-263
2. Yewdell J, Bennink J: **Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses.** *Annu Rev Immunol* 1999, **17:**51-81
3. Rotzschke O, Falk K, Stevanovic S, Jung G, Rammensee H: **Peptide motifs of closely related HLA class I molecules encompass substantial differences.** *European Journal of Immunology* 1992, **22:**2453-2456
4. Rammensee H-G, Friede T, Stevanovic S: **MHC ligands and peptide motifs: first listing.** *Immunogenetics* 1995, **41:**962-965
5. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins.** *Proc Natl Acad Sci U S A* 1987, **84:**4355-4358
6. Rammensee H-G, Bachman J, Philipp N, Emmerich N, Bachor OA, Stevanovic S: **SYFPEITHI: a database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50:**213-219
7. Parker KC, Bednarek MA, Coligan JE: **Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains.** *J Immunol* 1994, **152:**163-175
8. Gulukota K, Sidney J, Sette A, DeLisi C: **Two complementary methods for predicting peptides binding major histocompatibility complex molecules.** *J Mol Biol* 1997, **267:**1258-1267
9. Baldi P, Brunak S: **Bioinformatics, the machine learning approach.** *MIT Press Cambridge Massachusetts, London England* 1998
10. Honeyman M, Brusic V, Stone N, Harrison L: **Neural network-based prediction of candidate t-cell epitopes.** *Nature Biotechnology* 1998, **16:**966-969
11. Mamitsuka H: **MHC molecules using supervised learning of hidden Markov models.** *Proteins: Structure, Function and Genetics* 1998, **33:**460-474
12. Brusic V, Harrison L: **Prediction of MHC binding peptides using artificial neural networks, Complex Systems:** *In Complex Systems: Mechanism of Adaptation (Edited by: Stonier RJ, Yu XS) IOS Press, Amsterdam, The Netherlands/OHMSHA Tokyo* 1994, 253-260
13. Schueler-Furman O, Altuvia Y, Sette A: **Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles.** *Protein Science* 2000, **9:**1838-1846
14. Brown M, Grundy WN, Lin D, Cristianini N, Sugnet CW, S T, Ares M Jr, Haussler D: **Genetics knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97:**262-267
15. Cristianini N, Shawe-Taylor J: **Support vector machines and other kernel-based learning methods.** *Cambridge University Press, Cambridge England The Edinburgh Building, Cambridge, CB2 2RU, UK* 2000

16. Ding C, Dubchak I: **Multi-class protein fold recognition using support vector machines and neural networks.** *Bioinformatics* 2001, **17:**349-358
17. Joachims T: **Making large-Scale SVM Learning Practical.** *In: Advances in kernel methods – support vector learning (Edited by: B Schölkopf and C Burges and A Smola) MIT Press, Cambridge Massachusetts, London England* 1999
18. Vapnik VN: **The Nature of Statistical Learning Theory.** *Wiley New York* 1998
19. Rost B, Sander C: **Prediction of protein secondary structure structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-599
20. Matthews B: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta* 1975, **405:**442-451
21. Brusic V, Rudy G, Harrsison LC: **MHCPEP, a database of MHC-binding peptides: update 1997.** *Nucleic Acids Research* 1998, **26:**368-171
22. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project.** *Nucleic Acids Research* 2002, **30:**38-41
23. Lindahl E, Elofsson A: **Identification of related proteins on family, superfamily and fold level.** *J Mol Biol* 2000, **295:**613-625

Publish with **BioMed** Central and every scientist can read your work free of charge

"BioMedcentral will be the most significant development for disseminating the results of biomedical research in our lifetime."

Paul Nurse, Director-General, Imperial Cancer Research Fund

Publish with **BM**C and your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours - you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/manuscript/

BioMedcentral.com

editorial@biomedcentral.com